Abstract—A driver’s gaze is critical for determining their attention, state, situational awareness, and readiness to take over control from partially automated vehicles. Estimating the gaze direction is the most obvious way to gauge a driver’s state under ideal conditions when limited to using non-intrusive imaging sensors. Unfortunately, the vehicular environment introduces a variety of challenges that are usually unaccounted for - harsh illumination, nighttime conditions, and reflective eyeglasses. Relying on head pose alone under such conditions can prove to be unreliable and erroneous. In this study, we offer solutions to address these problems encountered in the real world. To solve issues with lighting, we demonstrate that using an infrared camera with suitable equalization and normalization suffices. To handle eyeglasses and their corresponding artifacts, we adopt image-to-image translation using generative adversarial networks to pre-process images prior to gaze estimation. Our proposed Gaze Preserving CycleGAN (GPCycleGAN) is trained to preserve the driver’s gaze while removing potential eyeglasses from face images. GPCycleGAN is based on the well-known CycleGAN approach - with the addition of a gaze classifier and a gaze consistency loss for additional supervision. Our approach exhibits improved performance, interpretability, robustness and superior qualitative results on challenging real-world datasets.

I. INTRODUCTION

Driver safety and accident risk are highly determined by a driver’s attention levels, especially visual attention, e.g. a driver’s gaze, for both conventional and self-driving vehicles [1]–[5]. It is crucial to have the ability to monitor the driver’s gaze in real-time and use gaze information to efficiently enhance vehicle safety for intelligent vehicles. Previous work has proven that visual information from imaging sensors and corresponding learning algorithms are effective in determining driver gaze zones [6]–[9] and driver activity in general [10]–[12]. However, the state-of-the-art for gaze estimation does not account for the complexities of the real world (for example - drivers wearing eyeglasses, harsh illumination, nighttime conditions, etc.). According to the Vision Impact Institute [13], [14], one in five drivers have a vision problem and most of them wear correction eyeglasses while driving. Additionally, according to the National Safety Council, 50% of all accidents happen while driving in the dark [15]. However, it is difficult to estimate the gaze with a high accuracy using RGB images acquired from dark environments. It is primarily because RGB cameras suffer from a lack of photons captured by imaging sensors, resulting in low signal-to-noise ratios [16].

In this study, we find that using infrared cameras can mitigate these lighting issues; however, for face images with eyewear, data pre-processing or algorithmic improvements are necessary to identify the driver’s gaze zones. Since previous works on gaze estimation [8] achieved decent results (under ideal conditions) by using convolutional neural networks (CNNs), we adopt their basic methodology and develop novel pre-processing approaches that can improve gaze estimation in these demanding conditions.

Generative Adversarial Networks (GANs) [17] have shown promising results on a variety of computer vision tasks like image super-resolution [18], realistic face images generation [19], style transfer, image editing [20], [21], etc. There are two sub-models in GAN architecture, a generator and a discriminator. The generator updates itself to synthesize images that are indistinguishable from real images by learning the data distribution. The discriminator’s task is to differentiate between real and synthesized images. Both of them are trained together in an adversarial manner until an equilibrium is attained. More recent work on GANs has focused on addressing key issues like training on large datasets, training larger models, improving stability during training, preserving finer details, etc. In this study, we build on one such model called the CycleGAN [21], which achieved the state-of-the-art on image-to-image translation by using unpaired images from the source and target distributions.

Inspired by the CycleGAN architecture, we propose the Gaze Preserving CycleGAN (GPCycleGAN) which makes use of an additional gaze consistency loss. GPCycleGAN has the following advantages compared to other gaze estimation methods and glass removal techniques: First, GPCycleGAN

Fig. 1: The real world introduces variability and complexity that driver gaze estimation systems usually ignore. Some examples include - use of eyeglasses, harsh illumination, nighttime data etc.
preserves the gaze and allows for more accurate gaze estimation on images with eyeglasses. Second, unlike previous works [25], [26] on eyeglass removal, there is no need for paired images to train GPCycleGAN. Third, it works in different environments, lighting conditions, eyeglass types, and with significant variations of the head pose.

The four main contributions of this paper are: a) An in-depth analysis of traditional gaze estimation under different conditions (e.g., with and without eyeglasses, daytime, nighttime, etc.), and its shortcomings, b) The GPCycleGAN model for eyeglass removal specifically optimized for the gaze classification task, c) Experimental analyses to illustrate how the GPCycleGAN model improves the accuracy, interpretability and robustness over a variety of baseline models, and d) A naturalistic driving dataset with labeled gaze zones that includes both IR and RGB images. This work is an extension of our research presented in [29]. In this study, we extend and refine our previous approach with new losses and training schemes, update our previously reported quantitative numbers, and provide new qualitative analysis.

II. RELATED RESEARCH

Recent related studies are listed in Table I with 2 sub-tables that represent two research areas - gaze estimation and eye-glass removal. We mainly focus on vision-based approaches post 2016 for gaze estimation. Please refer to the following studies for earlier work: Vora et al. [8] for gaze estimation; Kar et al. [30] for eye-gaze estimation systems, algorithms, and performance evaluation methods in consumer platforms; a survey on driver behavior analysis for safe driving by Kaplan et al. [31]; a review on driver inattention monitoring systems by Dong et al. [32]; and a survey on head pose estimation by Murphy et al. [33]. For the second research area, we were unable to find published studies focused on eyeglass removal for drivers’ face images. We instead provide discussions on general-purpose eyeglass removal papers.

### A. Gaze Estimation

Gaze estimation studies can be categorized in the following ways. First, by the type of model used, they can be divided into convolutional neural networks [8], statistical learning models [34], or geometric approaches [7]. Second, methods can be distinguished by the cues they consider, i.e. studies that only use head pose [6] versus ones that use both head pose and eye information [8], [9], [35], [36]. Third, research can be categorized by the conditions they capture. For instance, studies with limited illumination changes [8] versus research with multiple environments and the variations that come with it [9]. Lastly, studies can be separated by the sensors they use, including RGB cameras [8], IR cameras [37], NIR cameras [22], multiple cameras [38], and wearable sensors [39]. However, only a few studies emphasize the problems associated with driving with eyeglasses or eye-wear. Tawari et al. [38] and Lee et al. [40] mention the unreliability of gaze estimation for drivers with glasses, and propose methods that only rely on head pose as fallback solutions. Naqvi et al. [22] combine head pose estimation and pupil detection to determine the eye gaze, but their method only works under ideal capture conditions. Jah and Busso [6] use only head pose in their method. Wang et al. [23] combine depth images of the head and RGB images of eyes, but gaze estimation for eye images

### TABLE I: Related research

(a) Selected research on gaze estimation in vehicular environments

<table>
<thead>
<tr>
<th>Study</th>
<th>Objective</th>
<th>Sensor</th>
<th>Features</th>
<th>Capture conditions</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vora et al. [8]</td>
<td>Gaze zone classification using CNNs</td>
<td>1 RGB camera</td>
<td>HP &amp; gaze</td>
<td>Daytime, w/o eyeglasses</td>
<td>CNN</td>
</tr>
<tr>
<td>Martin et al. [7]</td>
<td>Estimating gaze dynamics, glance duration and frequency</td>
<td>1 RGB camera</td>
<td>HP &amp; gaze</td>
<td>Daytime, w/o eyeglasses</td>
<td>CNN &amp; geometry</td>
</tr>
<tr>
<td>Naqvi et al. [22]</td>
<td>Gaze zone detection using NIR camera and deep learning</td>
<td>1 NIR camera &amp; NIR LEDs</td>
<td>HP &amp; gaze</td>
<td>Daytime &amp; nighttime, w &amp; w/o eyeglasses</td>
<td>CNN</td>
</tr>
<tr>
<td>Yong et al. [9]</td>
<td>Gaze detection using dual NIR cameras and deep residual networks</td>
<td>2 NIR cameras &amp; NIR LEDs</td>
<td>HP &amp; gaze</td>
<td>Daytime &amp; nighttime, w &amp; w/o eyeglasses</td>
<td>CNN</td>
</tr>
<tr>
<td>Jha &amp; Busso [6]</td>
<td>Gaze region estimation using dense pixelwise predictions</td>
<td>1 RGB camera &amp; headband with AprilTags</td>
<td>HP</td>
<td>Daytime, w/ &amp; w/o eyeglasses</td>
<td>Dense Neural Networks</td>
</tr>
<tr>
<td>Wang et al. [23]</td>
<td>Continuous gaze estimation using RGB-D camera</td>
<td>1 RGB-D camera</td>
<td>HP &amp; gaze</td>
<td>Daytime, w/ &amp; w/o eyeglasses</td>
<td>Feature extraction &amp; k-NN</td>
</tr>
</tbody>
</table>

(b) Selected research on eyeglass removal and related topics

<table>
<thead>
<tr>
<th>Study</th>
<th>Objective</th>
<th>Methodology</th>
<th>Dataset</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhu et al. [21]</td>
<td>Unpaired image-to-image translation</td>
<td>Cycle-GAN³</td>
<td>Unpaired images</td>
<td>Uses unpaired images; performs well on style transfer</td>
<td>Not realistic for eyeglass removal; does not preserve gaze direction</td>
</tr>
<tr>
<td>Amodio et al. [24]</td>
<td>Unpaired image-to-image translation</td>
<td>TraVelGAN</td>
<td>Unpaired Images from multiple domains</td>
<td>Good performance on style transfer and eyeglass removal</td>
<td>Eyes are often swapped; gaze is not preserved</td>
</tr>
<tr>
<td>Li et al. [27]</td>
<td>Identity-aware transfereence of facial attributes</td>
<td>CNN</td>
<td>Images from web &amp; surveillance cameras</td>
<td>Single step, end-to-end method</td>
<td>Images need to be aligned; needs paired images; uses synthesized eyeglasses</td>
</tr>
<tr>
<td>Shen et al. [28]</td>
<td>Facial attribute manipulation</td>
<td>GAN</td>
<td>CelebA &amp; LFW</td>
<td>Flexibility to modify different facial attributes</td>
<td>Generated gaze is different; blurry outputs</td>
</tr>
</tbody>
</table>

¹ Head Pose  
² Near Infra-Red  
³ Generative Adversarial Networks
is unstable and only works with frontal images under ideal conditions. In [9], Yoon et al. collect a dataset comprising of images in daytime/nighttime, images with and without eyeglasses using two NIR cameras and NIR lights. Although they achieve good performance, they do not explicitly model the presence of eyeglasses in images. In this study, we show that such approaches tend not to generalize to different settings and usually overfit the training set.

B. Eyeglass Removal

In real-world cabin environment, large variations, including head pose, eyeglass type, environment conditions, and the presence of reflection artifacts, make eyeglass removal is a complex task. Existing eyeglass removal studies make use of statistical learning, principal component analysis (PCA), or deep learning, including GAN based methods. Statistical learning and PCA were the primary approaches prior to the advent of deep learning. These approaches do not require high computational power but have limitations on eyeglasses, environmental conditions, and head poses [41], [42]. On the other hand, methods using deep learning, such as ones proposed by Liang et al. [26] and Wang et al. [25] modified GANs for better results, but both of these methods need paired and aligned images for training. Such datasets are expensive and tedious to collect. Different GAN architectures proposed in [21], [24], and [43] do not require paired images, but their models fail to keep gaze information and do not perform well on non-frontal images. Models developed by Li et al. [27] and Shen et al. [28] are capable of changing multiple facial attributes but do not produce satisfying results on eyeglass removal. All the above methods on eyeglass removal have general or different purposes. However, none of these methods are constructed to preserve gaze of face images, and hence cannot necessarily be used out-of-the-box for gaze estimation.

III. DATASET

Since our primary goal is to design a gaze estimation system for the real-world, we prioritized using small form-factor infrared cameras with suitable real-time performance. We decided on an Intel RealSense IR camera and enclosed it in a custom 3D printed enclosure mounted next to the rearview mirror. To ensure a good compromise between larger fields-of-view and faster processing speeds, we settled on a capture resolution of \(640 \times 480\). Similar to Vora et al. [8], we divide the driver’s gaze into seven gaze zones: Eyes Closed/Lap, Forward, Left Mirror, Speedometer, Radio, Rearview, Right Mirror. Unlike previous studies, we also include gaze zones related to driver inattention or unsafe driving behavior. Our entire dataset comprises of thirteen subjects in different lighting conditions (daytime, nighttime and harsh lighting), wearing a variety of eyeglasses. For every gaze zone, participants were instructed to keep their gaze fixed while moving their heads within reasonable limits. In total, 336177 frames of images were captured, which we split into training, validation, and test sets with no overlap of subjects. Table II shows the distribution of images and subjects across different splits and capture conditions. Figure 2 depicts exemplar images from our dataset.

TABLE II: Dataset size (# of images, # of subjects) across different splits and capture conditions. There is no overlap of subjects between different splits to ensure cross-subject validation and testing.

<table>
<thead>
<tr>
<th>Capture conditions</th>
<th>Dataset split</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>daytime; w/o eyeglasses</td>
<td>(67151, 9)</td>
<td>(9908, 1)</td>
<td>(2758, 4)</td>
<td></td>
</tr>
<tr>
<td>nighttime; w/o eyeglasses</td>
<td>(59352, 9)</td>
<td>(8510, 1)</td>
<td>(2768, 4)</td>
<td></td>
</tr>
<tr>
<td>daytime; w/ eyeglasses</td>
<td>(34342, 9)</td>
<td>(9062, 1)</td>
<td>(3294, 4)</td>
<td></td>
</tr>
<tr>
<td>nighttime; w/ eyeglasses</td>
<td>(33189, 5)</td>
<td>(8103, 1)</td>
<td>(2897, 4)</td>
<td></td>
</tr>
<tr>
<td>Total (all conditions)</td>
<td>(203124, 9)</td>
<td>(35583, 1)</td>
<td>(11717, 4)</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2: Example images from our dataset under different capture conditions.

We ensure that the dataset is suitably diverse and challenging to represent the complexities observed in the real world.

IV. METHODOLOGY

As depicted in Fig. 3, the steps involved in our proposed gaze estimation pipeline are as follows: (a) landmark detection using OpenPose [44], (b) eye image cropping, resizing, and equalization, (c) gaze estimation. We use OpenPose for landmark detection because of its high accuracy under different conditions and fast inference speed. Based on the work by Vora et al. [8], we crop the eye region using the estimated landmarks as per their conclusion that the upper half of the face as an input produced the best results for downstream gaze classification. Next, we use adaptive histogram equalization to improve the contrast and resize the images to \(256 \times 256\) before feeding them to the gaze estimation models.

A. Issues with Lighting & Eyeglasses

To understand the impacts of different conditions on gaze estimation, we carry out an extensive experiment to analyze the performance of models trained with subsets of data, when they are tested on data from within and outside the training distribution. Table III shows validation accuracies of SqueezeNet-based gaze classifier models \(\text{Model 1} - \text{Model 9}\) (as proposed in [8]), each trained on data captured under different conditions \(\text{Model 5} - \text{Model 9}\). From the Table, we glean that model \(\text{Model 5}\) validated on data \(\text{Model 4}, \text{Model 6}, \text{and Model 2}\) produce similar accuracies, and model \(\text{Model 9}\) validated on data \(\text{Model 2}, \text{Model 6}, \text{and Model 1}\) also perform similarly. This demonstrates that the gaze classifier models work well
TABLE III: Validation accuracies for gaze models trained on data with different capture conditions

<table>
<thead>
<tr>
<th>Model &amp; training data used</th>
<th>Validation data</th>
<th>( \mathbb{G} ) daytime; w/o eyeglasses</th>
<th>( \mathbb{G} ) nighttime; w/o eyeglasses</th>
<th>( \mathbb{W} ) daytime; w/ eyeglasses</th>
<th>( \mathbb{W} ) nighttime; w/ eyeglasses</th>
<th>( \mathbb{W} ) w/o eyeglasses</th>
<th>( \mathbb{B} ) daytime</th>
<th>( \mathbb{B} ) nighttime</th>
<th>( \mathbb{B} ) all conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 daytime; w/o eyeglasses</td>
<td>81.0774</td>
<td>62.4657</td>
<td>45.4588</td>
<td>11.0689</td>
<td>72.567</td>
<td>29.7526</td>
<td>64.1875</td>
<td>38.1299</td>
<td>52.2757</td>
</tr>
<tr>
<td>2 nighttime; w/o eyeglasses</td>
<td>73.5272</td>
<td>87.2226</td>
<td>29.3782</td>
<td>25.6684</td>
<td>79.7801</td>
<td>27.6839</td>
<td>52.5923</td>
<td>58.0671</td>
<td>55.0941</td>
</tr>
<tr>
<td>3 daytime; w/ eyeglasses</td>
<td>60.4746</td>
<td>42.7325</td>
<td>70.6334</td>
<td>39.7423</td>
<td>52.3625</td>
<td>56.5256</td>
<td>65.2918</td>
<td>41.3162</td>
<td>54.3356</td>
</tr>
<tr>
<td>4 nighttime; w/ eyeglasses</td>
<td>51.4649</td>
<td>76.7639</td>
<td>45.3773</td>
<td>64.8151</td>
<td>63.0322</td>
<td>54.2544</td>
<td>48.5782</td>
<td>71.1043</td>
<td>58.8720</td>
</tr>
<tr>
<td>5 w/o eyeglasses</td>
<td>85.5508</td>
<td>88.7809</td>
<td>43.9101</td>
<td>20.6954</td>
<td>87.0276</td>
<td>33.3080</td>
<td>65.8025</td>
<td>61.5675</td>
<td>61.5675</td>
</tr>
<tr>
<td>6 w/ eyeglasses</td>
<td>71.3746</td>
<td>82.2239</td>
<td>69.5738</td>
<td>77.2754</td>
<td>76.3351</td>
<td>73.2966</td>
<td>70.5207</td>
<td>80.0932</td>
<td>74.8951</td>
</tr>
<tr>
<td>7 daytime</td>
<td>77.2131</td>
<td>74.7569</td>
<td>73.5095</td>
<td>39.9223</td>
<td>76.0901</td>
<td>58.1704</td>
<td>75.4569</td>
<td>58.2573</td>
<td>67.5971</td>
</tr>
<tr>
<td>8 nighttime</td>
<td>55.1192</td>
<td>83.7696</td>
<td>49.3596</td>
<td>61.3520</td>
<td>54.8365</td>
<td>52.3881</td>
<td>73.1514</td>
<td>61.8763</td>
<td>61.8763</td>
</tr>
<tr>
<td>9 all conditions</td>
<td>81.6969</td>
<td>83.3084</td>
<td>56.7885</td>
<td>66.4358</td>
<td>69.8857</td>
<td>75.3166</td>
<td>72.3675</td>
<td>75.3166</td>
<td>72.3675</td>
</tr>
</tbody>
</table>

Fig. 3: Overall processing pipeline for driver gaze zone estimation.

on daytime and nighttime data when trained on data containing both conditions. Thus, problems related to lighting can be effectively solved by training using IR images, appropriate normalization, and histogram equalization. We also learn that nighttime data is easier to model and results in better accuracies in general. Next, to analyze the models’ performance on data with and without glasses, we observe that the accuracy of model 5 validated on data \( \mathbb{G} \) (87.0276\%) is much higher than that of model 4 \( \mathbb{W} \) (73.5272\%). This implies that simply training with glasses \( \mathbb{W} \) is better than training without glasses \( \mathbb{G} \). We denote \( \mathbb{G} \) as sample images from domains \( \mathcal{X} \) and \( \mathcal{Y} \) respectively.

First, we consider a baseline gaze classifier model (Step 1 in Fig. 4(b)), for which we use input images from domain \( \mathcal{X} \). We use the SqueezeNet architecture as before and denote its output gaze zone probabilities as \( \{p_i\}_i \). Standard cross-entropy loss \( \mathcal{L}_{CE} \) is used for training the gaze classifier. \( \mathcal{L}_{CE} \) is defined as:

\[
\mathcal{L}_{CE} = - \sum_i z_i \log p_i, \tag{1}
\]

where \( i \) is the class index and \( z \) is the one-hot encoding of the ground truth class.

Next, we consider a CycleGAN-based eyeglass removal network, followed by a SqueezeNet-based gaze classifier (Step 2 in Fig. 4(b)). The CycleGAN-based eyeglass removal model learns the mapping \( G_{w/o} \) from \( \mathcal{Y} \) to \( \mathcal{X} \) (removing eyeglasses), denoted as \( G_{w/o}: \mathcal{Y} \rightarrow \mathcal{X} \). When trained, \( G_{w/o} \) generates images \( G_{w/o}(\cdot) \) using the input from either domain, and then passes them to the pre-trained gaze classifier to generate output probabilities \( \{p_i\}_i \). In a similar manner, the mapping for adding eyeglasses to \( X \in \mathcal{X} \) is \( G_{w/o}: \mathcal{X} \rightarrow \mathcal{Y} \), and it is trained simultaneously with \( G_{w/o} \). The discriminator \( D_w \) aims to distinguish between real images \( \mathcal{Y} \) and generated images \( G_{w/o}(\cdot) \), whereas the discriminator \( D_{w/o} \) aims to do so between real images \( \mathcal{X} \) and generated images \( G_{w/o}(\cdot) \).

The losses used in this model are: cycle consistency losses [21] (Eq. 2) that encourages cycle consistency between the real images and the reconstructed images; adversarial losses [17] (Eq. 3) for making the generated images indistinguishable from real images; and identity losses [45] (Eq. 4) to ensure identity mapping when the input belongs to the target domain.

\[
\mathcal{L}_{cy}(G_{w/o}, G_{w/o}) = \mathbb{E}_{X \sim p(X)} \left[ \|G_{w/o}(G_{w/o}(X)) - X\|_1 \right] \\
+ \mathbb{E}_{Y \sim p(Y)} \left[ \|G_{w/o}(G_{w/o}(Y)) - Y\|_1 \right]; \tag{2}
\]

\[
\mathcal{L}_{adv}(G_{w/o}, G_{w/o}, D_{w/o}, D_{w/o}) \\
= \mathbb{E}_{Y \sim p(Y)} \left[ \log D_{w/o}(Y) \right] \\
+ \mathbb{E}_{X \sim p(X)} \left[ \log (1 - D_{w/o}(G_{w/o}(X))) \right] \\
+ \mathbb{E}_{X \sim p(X)} \left[ \log D_{w/o}(X) \right] \\
+ \mathbb{E}_{Y \sim p(Y)} \left[ \log (1 - D_{w/o}(G_{w/o}(Y))) \right]; \tag{3}
\]

and,

\[
\mathcal{L}_{ident}(G_{w/o}, G_{w/o}) = \mathbb{E}_{Y \sim p(Y)} \left[ \|G_{w/o}(Y) - Y\|_1 \right] \\
+ \mathbb{E}_{X \sim p(X)} \left[ \|G_{w/o}(X) - X\|_1 \right]; \tag{4}
\]
where \( X \sim P(X) \) and \( Y \sim P(Y) \) are the image distributions.

The full objective for the CycleGAN model is the following:

\[
L_{total} = L_{adv} + \lambda_1 L_{cyc} + \lambda_2 L_{identity},
\]

(5)

where \( \lambda_1 \) is the weight of the cycle loss, and \( \lambda_2 \) is the weight of the identity loss.

As a final refinement step, we retrain the gaze classifier from step 1 using real images from \( X \) and fake images from \( G_{w/o}(\cdot) \) (Step 3 in Fig. 4(b)). Unlike step 1, we use a selective cross-entropy loss \( (L_{sel-CE}) \) to avoid overfitting to errors resulting from \( G_{w/o}(\cdot) \). \( L_{sel-CE} \) is defined as:

\[
L_{sel-CE} = \begin{cases} 
- \sum_i z_i \log p_i, & \text{if } \arg \max_i p_i = \arg \max_i z_i \\
0, & \text{otherwise}
\end{cases}
\]

(6)

where only the cross-entropy loss from correctly classified samples is backpropagated. This ensures that the gaze classifier is not trained to incorrectly classify fake images with a potentially different gaze zone than the original.

### C. Proposed GPCycleGAN-based Approach

Our proposed GPCycleGAN model with the gaze classifier (Fig. 4(c)) builds on the previous model, with the addition of a gaze consistency loss \( (L_{gaze}) \) in step 2. To preserve the gaze features during eyeglass removal, we use the trained gaze classifier from step 1 and compute the gaze consistency loss using its Class Activation Maps (CAMs). The proposed gaze consistency loss is defined as follows:

\[
L_{gaze}(G_{w/}, G_{w/o}) = \frac{1}{N} \sum_{i=1}^{N} ||T(A_i^{w/o,real}) - T(A_i^{w/o,rec})||_2,
\]

(7)

where \( N \) stands for the total number of classes (gaze zones), \( i \) is the class index, and \( \{A_i\}_{i=1}^{N} \) are the CAMs from the trained gaze classifier. Subscript \( w/o \) real denotes CAMs obtained from a real eye image crop without eyeglasses, and subscript \( w/o \) rec denotes CAMs corresponding to reconstructed images \( G_{w/o}(G_{w/}(\cdot)) \) using the same eye image crop. The transformation \( T(\cdot) \) applied to the CAMs is a simple sigmoid.
The full objective for the proposed GPCycleGAN model is:

\[ L_{\text{total}} = L_{\text{adv}} + \lambda_1 L_{\text{cyc}} + \lambda_2 L_{\text{identity}} + \lambda_3 L_{\text{gaze}}, \]  

(9)

where \( \lambda_3 \) is the weight for the gaze consistency loss.

### D. Implementation Details

The training and inference architectures are the same for the baseline gaze classification models, but different for the proposed GPCycleGAN model. Models A and B are identical in terms of their architectures, and only differ in their training setup and losses (i.e., the addition of a gaze consistency loss in model B). Both models consist of generators with 9 residual blocks and 70 × 70 PatchGANs [21] as discriminators. As illustrated in Fig. 4, training models A and B proceeds in 3 steps: Step 1 involves training a SqueezeNet gaze classifier using eye image crops from domain \( X \); Step 2 is for training the generator \( G_{w/o} \) in CycleGAN/GPCycleGAN; Step 3 is to fine-tune the gaze classifier from step 1 using generated images \( G_{w/o}(x, y) \). The inference for models (A - G) is a simple forward propagation through the gaze classifier to obtain the gaze zone probabilities. Models A and B have the same inference setup (Fig. 5(b)), where eye image crops are first passed to the generator \( G_{w/o} \) for eyeglass removal, after which they are fed to the gaze classifier which outputs the gaze zone probabilities.

We choose \( \tau = 0.01 \) in Eq. 8, and \( \lambda_1 = 10, \lambda_2 = 5, \lambda_3 = 1 \) in Eq. 5 and Eq. 9. We use a learning rate of 0.0004 with an Adam optimizer for training all SqueezeNet classifiers, a learning rate of 0.0002 for CycleGAN/GPCycleGAN in step 2, and a reduced learning rate of 0.0001 for the final finetuning step. The Gaze classifiers are trained for a total of 50 epochs, while the GANs are trained for 15 epochs - both with early stopping.

### V. Experiments & Analyses

#### A. Quantitative Metrics & Results

Table IV presents the cross-subject test accuracies for 6 different models. We use two metrics, namely the macro-average and micro-average accuracy, defined as follows:

Macro-average accuracy

\[ \text{Macro-average accuracy} = \frac{\sum_{i=1}^{N} (\text{True positives})_i}{(\text{Total population})}, \]

(10)

Micro-average accuracy

\[ \text{Micro-average accuracy} = \frac{\sum_{i=1}^{N} (\text{True positives})_i}{\sum_{i=1}^{N} (\text{Total population})}, \]

(11)

where \( N \) is the number of classes. Micro-average accuracy represents the overall percentage of correct predictions, while macro-average accuracy represents the average of all per-class accuracies. So, if the data is balanced, which means the number of images for every class is the same, micro-average accuracy will equal macro-average accuracy. In our case, there are more data in higher accuracy classes, so the micro-average accuracies are higher than macro-average accuracies. Thus, we mainly compare the macro-average accuracy since the dataset is not balanced.

First, we consider model 5 trained only on images without eyeglasses to illustrate the domain gap between images with and without eyeglasses. As can be seen, this model performs poorly on a test set containing images outside its training distribution. Next, model 9 represents the scenario where the eyeglasses are not explicitly modeled. Although it is trained on the entire training set, the resulting accuracies indicate a performance penalty, especially when tested on images with eyeglasses.

Next, we see that adding a pre-processing network to remove eyeglasses such as in models A and B produces slightly worse accuracies than model 9. This can be attributed to the fact that the gaze classifier has only been trained on half the training set i.e. images without eyeglasses. However, after fine-tuning using the entire training set (Step 3 in Fig. 4), the accuracies for both models A and B increase considerably. In conclusion, our proposed model B demonstrates significant improvement over both the baseline model 9 and the vanilla CycleGAN-based model A after fine-tuning. The above evidence implies the benefits of our proposed gaze consistency loss, and demonstrates that the generator resulting from the
GPCycleGAN model acts effectively as a pre-processing step for the downstream task of gaze estimation.

In addition to the test set accuracies, we also present the corresponding confusion matrices of all 6 models in Fig. 6. The error modes of our best performing model (Fig. 6f) can mostly be attributed to confusion between gaze zones close in physical space (for example, Forward versus Speedometer), occlusion of the eyes by the eyeglass frame and glare, and the inability to distinguish between looking downwards versus closed eyes.

B. Qualitative Results

To carry out a qualitative comparison between different GAN variants, we also show 10 examples of eyeglass removal on real images using CycleGAN and GPCycleGAN in Fig. 7. In columns i, iii, v, vi, and vii, GPCycleGAN not only removes the eyeglasses, but also removes the glare resulting from it; whereas CycleGAN perceives the glare as part of the sclera. Glare removal is essential for gaze estimation because the glare from glasses is a relatively common occurrence in the real world, and often occludes the eyes, making it harder for models to learn discriminative gaze features. In columns ii, iv, vi, and viii, the images generated from GPCycleGAN are realistic and preserve the gaze more accurately. Columns ii, iii, iv, vi, and vii show that our model does not only work with frontal face images but also performs well for a variety of head poses. Column ix is an example where both models perform well because the gaze is clear and not occluded by the frame or glare. The last column (x) depicts a failure case for both models. The models fail because the frame of the eyeglass is too thick, and both the frame and glare occlude the eye regions severely. These problems could potentially be solved by collecting more data with thicker eyeglass frames, increasing the image resolution, and/or by designing better GANs.

To test the viability of our approach on different data types, we repeated our prescribed training procedure using an RGB dataset collected for the same task. Fig. 8 depicts qualitative results for eyeglass removal using our RGB model. For this RGB dataset, we notice that the resulting images after eyeglass removal tend to be sharper, more realistic, and preserve more details. We believe this is because of the higher resolution and better signal-to-noise ratio of the RGB sensor used to capture the dataset. We provide links to download both IR and RGB datasets in our repository.

C. Analyzing the Class Activation Maps

Observing the Class Activation Maps (CAMs) produced by classifier models is a simple and effective way of interpreting CNNs and assessing their robustness. In this spirit, we visualize and contrast the CAMs from our proposed approach (model B) and the baseline solution (model 9). This visualization is presented in Fig. 9, where eye image crops from the test set are overlaid with the corresponding ground truth CAMs.
i.e. the CAMs corresponding to the correct gaze zone. In particular, we compare CAMs generated when raw eye image crops with eyeglasses are directly fed to a gaze classifier (as in model 9) with CAMs that are obtained by first removing eyeglasses from the eye crop images using a pre-processing network, then feeding this processed image to a fine-tuned gaze classifier (as in model B). The CAMs clearly indicate that aside from being superior in terms of quantitative metrics, our proposed model exhibits better interpretability, robustness, and more accurately localizes the eye regions. The baseline model - despite being superior in terms of quantitative metrics, our proposed model exhibits better interpretability, robustness, and more accurately localizes the eye regions. The baseline model - despite being superior in terms of quantitative metrics, and instead overfits to unrelated visual cues. This portends better generalization for our model in comparison to the baseline.

VI. CONCLUDING REMARKS

Reliable and robust gaze estimation on real-world data is essential yet hard to accomplish. A driver’s gaze is especially important in the age of partially automated vehicles as a cue for gauging driver state/readiness. In this study, we improved the robustness and generalization of gaze estimation on real-world data captured under extreme conditions, such as data with the presence of eyeglasses, harsh illumination, nighttime driving, significant variations of head poses, etc. For dealing with issues arising from bad lighting, we demonstrate that using an IR camera with suitable equalization/normalization suffices. For images that include eyeglasses, we present eyeglass removal as a pre-processing step using our proposed Gaze Preserving CycleGAN (GP-CycleGAN). The GP-CycleGAN enables us to train a generator that is capable of removing eyeglasses while retaining the gaze of the original image. This ensures accurate gaze zone classification by a downstream SqueezeNet model. We show that this combined model outperforms both the baseline approach and the vanilla CycleGAN + SqueezeNet model considerably. We also provide example results and illustrations to demonstrate the superiority of our model in terms of interpretability, robustness, and generality. Future work entails improving on the architectures of different components like generator, discriminator, and gaze classifier.

We would like to thank the Toyota Collaborative Safety Research Center (CSRC) for their generous and continued support. We would also like to thank our colleagues at the Laboratory for Intelligent and Safe Automobiles (LISA), UC San Diego for their useful inputs and help in collecting and labeling the dataset.

REFERENCES


Fig. 9: IR eye image crops with ground truth CAMs produced by the gaze classifier. Top row: raw eye image crops with ground truth CAMs produced by model \( \Theta \). Bottom row: same eye image crops after eyeglass removal and updated ground truth CAMs produced by model \( \Theta \).


Akshay Rangesh is a PhD candidate in electrical engineering from the University of California at San Diego (UCSD), with a focus on intelligent systems, robotics, and control. His research interests span computer vision and machine learning, with a focus on object detection and tracking, human activity recognition, and driver safety systems in general. He is also particularly interested in sensor fusion and multi-modal approaches for real time algorithms.

Bowen Zhang completed his Masters in Electrical Engineering from the University of California at San Diego (UCSD), with a focus on machine learning and data science. His research interests include human activity recognition, embedded systems and sensing, and human-computer interaction. He is currently a Ph.D. student at the University of California at Santa Barbara (UCSB).

Mohan Manubhai Trivedi Mohan M. Trivedi (Life Fellow IEEE, SPIE, and IAPR) is a Distinguished Professor of Electrical and Computer Engineering at the University of California, San Diego and the founding director of the Computer Vision and Robotics Research (CVRR, est. 1986) and the Laboratory for Intelligent and Safe Automobiles (LISA, est. 2001). His research is in intelligent vehicles, intelligent transportation systems (ITS), autonomous driving, Driver Assistance Systems, active safety, human-robot interactivity, machine vision areas. UCSD LISA was awarded the IEEE ITSS Lead Institution Award. The CVRR-LISA team members have won numerous awards, including best dissertations, best papers, and best posters at major international conferences. Trivedi has served as the editor-in-chief of the Machine Vision and Applications, Senior Editor of the IEEE Trans on IV and ITSC, as well as Chairman of the Robotics Technical Committee of IEEE Computer Society and Board of Governors of the IEEE ITSS and IEEE SMC societies. Trivedi regularly serves as a consultant to Government agencies (including the NAE), Industry, and academic institutes in the USA and abroad.